图书情报工作
第 64 卷 第 14 期　2020 年 7 月

机器学习在术语抽取研究中的文献计量分析*

■ 邱科达　马建玲
中国科学院兰州文献情报中心　兰州 730000　中国科学院西北生态环境资源研究院　兰州 730000
中国科学院大学经济与管理学院图书情报与档案管理系　北京 100049

摘　要：[目的/意义]梳理和总结基于机器学习的自动术语抽取的相关研究，为领域相关人员提供参考。[方法/过程]在 CNKI 和 EndNote 的分析工具基础上，应用文献计量对主题的年度趋势和核心机构进行宏观分析，然后从抽取技术方法、数据集和评价以及应用 3 个方面进行主题内容分析。[结果/结论] 近些年，术语抽取研究取得了很大的进步，是知识系统、自然语言处理、情报分析等领域的基础工作。随着自然语言处理领域的迅猛发展，抽取技术开始朝着深度学习方向发展，但术语抽取的基础理论体系还有待完善，如评价指标、语料选取和效果评价方法。

关键词：术语抽取　机器学习　知识组织　文献计量
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.14.010

1　引言

随着语义网的发展，知识内容变革的范围逐渐扩大，步伐不断加快，知识载体的多渠道、多格式、关联数据化等异构现象已成为常态，用户群体更加渴望对知识内容的有效获取。语义网以知识组织为基础，试图实现知识之间的语义互联互通。其中，分类法、知识本体和知识图谱等知识密集型组织系统在语义网中扮演着重要的角色，能够揭示知识单元之间的内涵语义、挖掘知识外延关联，实现数据知识化、知识有序化以及知识服务化，最终让知识得到有效利用、传播、共享和增值。

术语是特定专业领域中概念的语言指称[1]。知识密集型系统需要大量准确、规范的术语来实现知识的表达、挖掘和可视化，是解决“信息和知识孤岛问题”的最佳方法。术语抽取（或术语识别）是从特殊领域文本中获得表示领域概念术语的过程，传统的术语抽取主要依赖专家知识来手工制定规则以进行术语的识别与抽取，存在规则维护扩展困难、应用范围有限、可移植性差等问题。在大数据时代，可获取的领域文本、词汇和概念等不断增长，手动构建、维护修订、索引和描述领域核心术语变成了一项劳动密集型任务，因此自动术语抽取（automatic term extraction，ATE）成为了领域术语自动获取研究的首要任务和基础工作。

自动术语抽取仍然是一个尚未解决的问题[2]，多年来学者们已经开发出了新的方法以满足工业、政府档案馆和数字图书馆对不断增长的专业文档自动归类标引的需求。这些方法通常结合了语言规则和统计信息，先利用语言处理器来提取候选术语（例如名词、名词短语或 n-gram），然后应用统计方法通过局部和全局收集的特征对候选者评分，最后对评分后的候选词进行排名，以供后续选择和筛选。现有方法已取得不错的抽取效果，但还存在两个局限：①众所周知，不可能针对任何领域开发一种不切实际的“一刀切”方法。研究[3]表明根据领域和数据集的不同，性能最佳的 ATE 方法总是会发生变化，并且不同方法获得的精度可能会显著不同。②目前最先进的技术通常利用词频之类的统计特征来对候选词进行评分，忽略了语义相关性的作用。

近些年，机器学习在术语抽取领域快速发展，学术价值和应用前景不断被探索和挖掘，从理论、模型、算法到实际应用都涌现出了很多优秀成果。通过自动学

* 本文系国家自然科学基金面上项目“气候变化科学成果集成研究范式及其实现平台研究”（项目编号：41671535）和中国科学院文献情报能力建设专项“开放学术资源体系”（项目编号：Y7ZG081001）研究成果之一。
作者简介：邱科达（ORCID：0000-0002-2826-8899），硕士研究生；马建玲（0000-0003-4933-5904），信息系统部副主任，研究馆员，硕士生导师，通讯作者：E-mail：majl@lzb.ac.cn。
收稿日期：2019-08-23　修回日期：2020-04-16　本文起止页码：94-103　本文责任编辑：易飞

邱科达，马建玲.机器学习在术语抽取研究中的文献计量分析[J].图书情报工作,2020,64(14):94-103.

习特征和截止点的最佳组合,机器学习能够有效地组合特征,具有广泛的适用性。面对领域数据的与日俱增和复杂多样,作为机器学习分支的深度学习可能会是更适合的选择。术语抽取是一项复杂而困难的工作,机器学习算法的融入进一步提升了术语抽取效果和质量,但通用性以及效果还有上升空间。基于术语抽取与机器学习的联系,本文将对该领域术语抽取方面的研究进展、应用情况等进行统计计量分析,为本领域相关人员提供参考。

2 相关文献定量分析

自动术语抽取研究已经有20多年的历程,早在20世纪90年代,国外就研究出了一批具有可操作性的术语抽取系统,如FASTER系统以及Terms系统等[4],服务于信息组织与检索、文本处理以及领域知识发现、组织与应用等方面。中文术语抽取研究起步较晚,主要是在国外研究基础上,结合汉语特点实现对已有方法的改进。目前,国内外有很多术语服务平台和工具,如中国科学技术信息研究所的《汉语主题词表》服务系统、全国科学技术名词审定委员会的术语知识服务平台Termonline、中国知网的知识元检索、OCLC术语服务以及Sketch Engine等。

文献信息资源日益增长,为了更加全面地获取与术语抽取研究相关的文献,本文以CNKI和维普为主要的中文检索平台,选取“术语抽取、术语识别、术语获取”为一级主题词进行主题检索,然后在检索结果中以“机器学习、深度学习、神经网络、监督学习、半监督学习、无监督学习、条件随机场、支持向量机、最大熵、隐马尔可夫模型”为检索词进行二次检索,分别得到290篇和126篇文献。再根据文献题目、摘要以及关键词,筛选并去重后得到96篇相关文献。对于外文文献,使用Web of Science核心合集数据库的高级检索功能,结合主题词构造如下检索式:“TS = (“term extraction” or “term recognition” or “terminology extraction” or “terminology recognition” or “term identification” or “terminology identification”) and ALL = (“machine learning” or “deep learning” or “neural network” or “conditional random fields” of “Support Vector Machine” or “supervised learning” or “unsupervised learning” or “Maximum Entropy” or “Hidden Markov Model”)”。在检索得到的79篇文献中,筛选后得到相关文献73篇。

考虑到数据的可获得性和研究成果的质量,本文合并收集到的国内外文献,采用文献计量和内容分析法对研究问题进行分析和论述。首先利用CNKI的数据分析功能以及EndNote的Subject Bibliography分析功能来获取文献的相关统计数据;然后用Excel对总的169篇文献进行年度趋势分析和核心研究机构分析,达到对相关研究的宏观认识;之后深入到文献的内容,结合统计数据,从抽取技术、数据集和评价以及应用方面对基于机器学习的术语抽取研究进行主题内容分析。

2.1 年度趋势分析

图1展示了各年度的发文情况,揭示了该研究的各个发展阶段。从文献发表时间来看,基于机器学习的术语抽取研究大约始于20世纪90年代,P. Marshall等[5]最早发表了会议论文*Working towards connectionist modeling of term formation*,研究是对术语识别的连接主义[6]方法研究的延续,提出了一种利用竞争性网络技术(赢家通吃算法,winner-take-all)来进行自动术语识别的方法。中文研究中,陈文亮等[7]于2003年应用Bootstrapping的机器学习算法,从大规模无标注真实语料中自动抽取领域词汇。基于机器学习的术语抽取研究从2007年开始进入上升期,随后在2009到2013年之间进入了一个稳定期,平均年发文量约10篇,这一时期,条件随机场、支持向量机、领域本体、专利分析等已经开始成为了术语抽取研究的关键词。在2012年左右,深度学习和大数据都进入到了快速发展阶段,推动了命名实体识别、关键词抽取、关系抽取等信息抽取研究的发展。作为信息抽取研究方向之一的术语抽取研究也受到影响,从2014年开始进入到了另一个上升期。近三年发表的文献主要探索神经网络或深度学习在术语抽取研究中的应用,这在一定程度上印证了图1中的上升趋势。

2.2 核心研究机构分析

研究机构是进行一项或多项学科研究的专门性组织,研究机构的计量分析可以揭示该研究方向的机构分布,辅助研究者找到学术跟踪信息源。通过对中英文169篇文献的机构进行统计和去重,图2列出了发文量大于3篇的14个主要机构(不排除合著的情况)。沈阳航空工业学院是沈阳航空航天大学的旧称,统一为沈阳航空航天大学,发文量最高,为12篇文献;其次为南京大学,发表了10篇。发文量最高的国外机构是曼彻斯特大学(Univiersity of Manchester),也是唯一一个发文量大于3的国外机构。通过对国外相关研究的调研发现,国外术语抽取研究更多集中在应用方面,嵌入到了本体、知识图谱、知识系统、自然语言处理等领

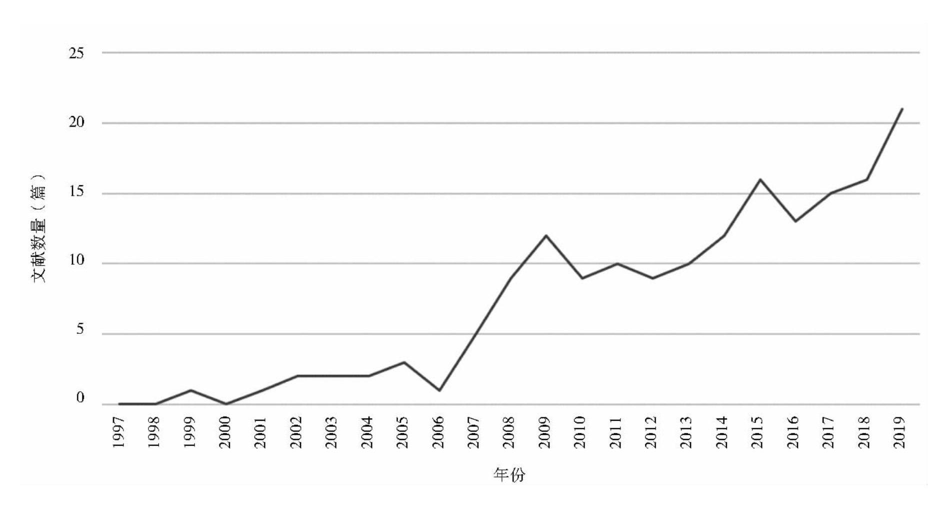


图 1　论文发文量年度趋势

域的研究中。相比较而言，国内机构发文主题较为单一、集中。中文术语抽取的研究更关注如何提高已有技术方法在中文上的表现，集中在医学和专利领域，较少开展在不同领域的应用研究。

通过统计可知，14 个机构一共发表了 70 篇，占全部文章总数的 42%，与其他研究单位相比，图 2 中所示机构有着较明显的优势，尤其是沈阳航空航天大学和南京大学，可以作为相关研究者今后的重点关注机构。同时，笔者发现这 14 个机构占整体比例非常小，可见术语抽取领域还缺乏高产的、杰出的研究机构。

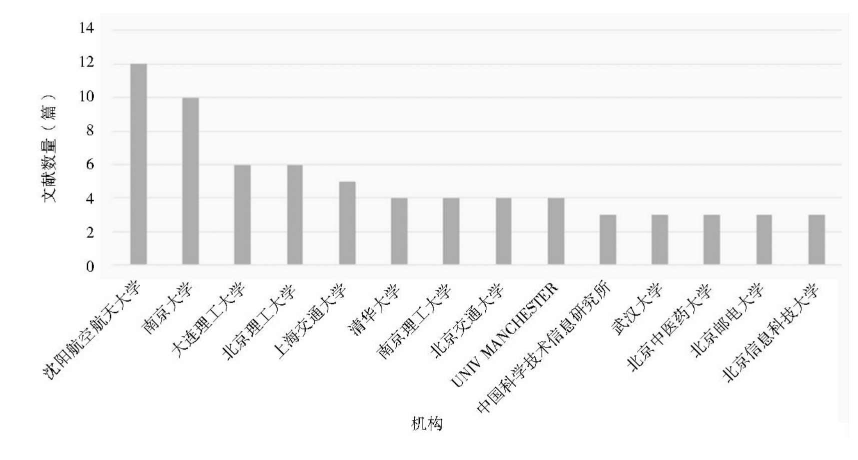


图 2　机构发文量分布

3　主题分析

主题分析能够反映领域的研究水平和总体状况，揭示领域的研究现状、热点及发展趋势。结合上文的宏观分析结果，笔者以研究对象的内容为切入点，对术语抽取领域相关论文的内容进行了主题分析，主要从以下 3 个方面对基于机器学习的术语抽取研究进行梳理：抽取技术方法、数据集和评估以及应用。

3.1　抽取技术方法分析

传统的术语抽取方法有基于语言学的方法、基于统计的方法、多策略混合的方法。基于语言学的方法[8-9]常依赖人工进行浅层语法分析或领域词典构建规则进行术语抽取，依赖特定语言、领域的词典、标注数据、知识库等先决资源条件，存在语言规则维护更新困难、可扩展性、可移植性差等问题，尤其对一些未登录词识别较差，准确率和召回率低，无法大规模地应用于术语抽取。基于统计的方法[10-11]利用术语在领域文本语料中的分布统计属性，将满足阈值或条件的字符串序列识别为领域术语，常用的指标 TF-IDF、信息熵、互信息、对数似然等，存在计算量大、容易遗漏低频词、忽略或缺乏上下文语义分析等问题。不同的方法也可以互相集成，结合多种策略来提升抽取性能。在统计和语言方法的基础上，K. T. Frantzi 等[12]提出 C-value/NC-value 方法，较早地开始了对混合策略的研究，其基本思想是先用规则模板得到候选术语集，然后使用统计特征来进行过滤；另外，周浪等[13]结合子串

归并、搭配检验和领域相关度计算技术来完善了中文词组型术语抽取系统的性能。

传统的术语抽取方法能够在特定语料上获得不错的表现，在多源异构数据和领域交叉的背景下，却愈发显得笨重。为了突破上述局限，随着机器学习和自然语言处理技术的推动，之后大多数研究开始将命名实体识别的方法迁移到术语抽取研究中，主要采用半监督和监督方式混合的机器学习算法及其变体等，聚焦于从领域文本中半自动或自动地获得领域依赖的属性、专门的文本特征、上下文语义信息等，以解决上述问题[14]。

为了分析机器学习在自动术语抽取中的技术方法，本文统计了相关文献的关键词。关键词是论文研究内容的高度提炼，笔者主要进行了同义词合并，如“CRF”和“条件随机场”；除去一些无益于主题研究的高频关键词，如“领域术语”“分词”“研究方法”；去除线条过粗的主题词，如“术语抽取”“机器学习”。表 1 展示了 169 篇文献中出现频率大于等于 4 的中英文关键词。综合中英文关键词来看，“条件随机场”“支持向量机”“神经网络”“深度学习”技术方法出现频率都在 4 次以上，其中条件随机场总共出现 53 次，“深度学习”出现 23 次，是近些年应用较多的术语抽取技术方法。根据关键词的分布，笔者从抽取技术自身特点以及发展时间线两个角度出发，将抽取方法分为统计机器学习方法和深度神经网络方法。

表 1　高频关键词

中文		英文	
关键词	词频	关键词	词频
条件随机场	49	natural language processing	12
本体(本体学习)	19	deep learning	11
深度学习	12	sentiment analysis	7
术语关系抽取(提取)	12	ontology (ontology learning)	7
神经网络	9	information extraction	5
信息抽取	8	neural network	5
专利术语	8	opinion mining	4
命名实体识别	5	conditional random fields	4
序列标注	5	aspect term extraction	4
自然语言处理	4	text mining	4
支持向量机	4		
文本挖掘	4		

3.1.1　统计机器学习方法

随着机器学习在自然语言处理领域的快速发展，术语抽取研究也逐渐转向了火热的机器学习阵营。基于统计机器学习的术语抽取研究可以总结为 3 个方向：模型选择、方法改进和多策略融合。

(1)模型选择。基于机器学习的术语抽取方法归根结底都是分类的方法，可以分为两种思路，一种是先识别出术语的边界，然后再进行分类；另一种是转化为序列标注问题。

分类模型是监督学习中一个典型的统计学习模型，主要是从已标注的训练数据中学习分类模型的权值、参数，用以预测新样本的类别。P. Lopez 等[15]为了抽取科技文档中的术语，比较分析了多种分类方法：决策树、支持向量机和多层感知机；赵欣[16]利用大量已有的术语，借助最大熵模型训练了术语分类器；M. Shirakawa 等[17]提出了一种扩展的朴素贝叶斯模型来实现文本中关键术语的抽取，以此实现对嘈杂短文本的分类；W. Zeng 等[18]使用 SVM 对新能源汽车领域的专利和文献数据进行术语抽取，实验结果证实了机器学习在术语抽取中的有效性。表 2 总结了已用于术语抽取的分类算法的功能和特点，这些算法还在文本分类、语音识别、图像理解等自然语言处理领域中取得了巨大成功。

序列标注模型能够解决自然语言中常见的问题，包括词性标注、命名实体识别、词义角色标注等。不同于一般的分类模型，序列标注模型将文本看作一个序列，利用 BIO、BIEO、BMEO 等标注方法进行术语的识别，是目前非常有效的方法。表 3 介绍了在术语抽取领域中常用的两种序列标注模型，隐马尔可夫模型(hidden markov model，HMM)和条件随机场(conditional random field，CRF)。H. S. Pan 等[19]针对汉语词库构建问题，提出了利用隐马尔可夫模型从学术文献中提取新术语的方法；岑咏华等[20]采用隐马尔可夫模型对计算机领域语料进行学习训练，F 值达到了 89.75%。相较于 HMM，CRF 更具优势，能够避免标记偏置问题，章成志[21]在条件随机场的基础上，提出了一种基于一体化策略的术语抽取方法；D. Zheng 等[22]把术语的离散特征作为 CRF 模板的属性，从单词本身、单词在组合型术语中的位置、文本的语义信息、信息熵和 TF/IDF 等多个角度调整特征模板，在领域术语识别任务中取得了不错效果。

(2)方法改进。机器学习在术语抽取研究上获得了迅速成功，为了设计出性能更好的术语抽取方法，研究者们对已有模型进行改进，提高了识别效果和计算效率，如 Q. Zhan 等[23]提出的层叠条件随机场模型。改进研究在中文术语抽取研究中较为多见，因为经典的模型大多面向英文，不能直接应用于汉语，根据汉语

表 2 基于分类的算法

算法	功能	特点
决策树	分类与回归方法（CART、ID3、C4.5、随机森林等）	复杂度小，速度快且抗噪能力强，可伸缩性好，既可用于小数据集，也可用于海量数据；在学习过程中使用者不需要了解背景知识，分类原理简单易懂
朴素贝叶斯	文本分类（VSM）	生成模型，原理简单，计算快；但属性间条件独立的假设太严苛，使得朴素贝叶斯的准确率受到影响
最大熵模型	文本分类（分类器）	特征选择灵活，模型可以应用在不同领域，可移植性强，但存在时空开销大、数据稀疏等问题
支持向量机	经典分类算法	可以解决线性不可分和多分类情况，计算复杂度仅取决于少量支持向量；但不支持类别型数据，难以在大规模数据上训练

表 3 基于序列标注的算法

算法	功能	特点
隐马尔可夫模型（HMM）	生成式模型	有向图模型，算法成熟，效果好，且易于训练，但只考虑了当前状态和观察对象，具有严格的独立性假设
条件随机场（CRF）	判别式模型	能够集成多个特征，克服了 HMM 的独立性假设，避免了标记偏置问题，但训练代价大，复杂度高

调整和优化经典模型，能够更有效地识别汉语文本中的术语。

统计机器学习方法的一个问题是依赖领域特定的特征工程。为了提高算法的精度，需要作为基础的专家知识（经验）和“运气”，即人工选取并获得最优特征的过程随机且不可控，因而难以大规模流行起来。因此，另一种提高术语抽取效果的思路是选择更好的特征表示[24]。术语抽取任务中，常用的特征包括形态、词汇和句法信息，形态特征有词形、前后缀等，词汇和句法特征包括词长、词性、浅层句法分析、依存句法分析等。考虑到汉语的特殊性，除了词汇层面的特征外，诸如偏旁部首、笔画等汉字层面的特征也被用来辅助提高术语抽取的效果。与此同时，各种外部知识如词典、维基百科、同义词林、HowNet、CN-Probase 等也可以提高识别性能。

（3）混合策略。混合策略方法能够有效减少模型的计算复杂度，并充分利用上下文语义信息进行领域文本分析，在一定程度上提高了识别的表现效果。C. Y. Chi 等[25]将基于独热编码的布朗聚类和隐马尔可夫模型结合起来实现对未标记语料库的无监督学习；另外，黄菡等[26]提出了结合主动学习的条件随机场模型，通过迭代的方式不断提高分类器的效率，准确率和召回率可达 90% 以上。

3.1.2 深度神经网络方法

2012 年以来，深度神经网络的激增和深度学习的发展，在语音识别、图像识别和计算机视觉等方面取得了丰硕的成果。尤其是基于词嵌入的语义表示方法，如 Word2Vec、fasttext、Glove、ELMo、BERT、XLNET 等，一方面解决了高维向量空间带来的数据稀疏问题，另一方面可以利用词嵌入从异构的文本中获取包含丰富语义信息的特征表示，为术语抽取这种带有领域性的序列标注问题，注入了强劲的发展动力。深度学习的优点是能够使用各种深度神经网络模型或算法从领域文本中自动学习特征，避免了繁重且耗时的特征工程，且学习特征的过程是人工、领域、语言非依赖性的，因而可移植、可重用、可扩展性强[14]。

为了解决现有机器学习方法中对特征工程的过度依赖和复杂问题泛化性能差等问题，近年来，一些研究开始探索基于深度神经网络方法的术语抽取。R. Chalapathy 等[27]发现传统的机器学习方法严重依赖人工特征和特定领域资源，提出使用 BLSTM-CRF 模型从临床数据中抽取医疗概念，取得了比 HMM、CRF 等 ATE 算法更优的结果。R. Wang 等[28]介绍了一种使用两个深度学习分类器进行术语抽取的弱监督自举方法，有效缓解了手工特征选择和标记数据缺乏的问题。

随着深度学习的不断发展，研究者们提出了一些优化机制。注意力机制实质是模拟人脑在特定时刻会将注意力集中在特定关键事物而忽略其他非关键事物的专注特性。马建红等[29]提出了基于 Attention 机制的 BLSTM-CRF 的领域术语抽取模型，准确率达到了 86%。迁移学习是从相关领域中迁移标注数据或者知识结构、完成或改进目标领域或任务的学习效果。刘宇飞等[30]引入深度迁移学习的思想，运用 BiLSTM 模型实现跨领域迁移，有效识别了技术术语，解决了专利文献少标注的问题。领域知识对于领域特定语料库中的术语抽取至关重要，很难从有限的语料库中获取知识。利用从诸如维基百科、百度百科等知识库中得出的领域事实，通过远程监督来学习术语特征，可以实现比现有方法更广的覆盖范围[31]。

当前术语抽取领域所应用的深度学习方法是在结合领域特点的基础上从命名实体识别研究中移植过来的，因此同样面临着缺乏大量规范文本、标注语料、基

础词库等领域资源条件的问题。从取得的研究结果来看,抽取精度有了较大提升,但尚未达到理想峰值。在深度学习技术基础上,如何提高抽取效率以及更有效地利用有限标注数据是术语抽取领域值得研究的方向,如在领域语料上利用预训练模型(BERT、XLNET)进行微调(fine-tune)。

3.2 数据集和评估分析

3.2.1 数据集分析

自动术语抽取是一个富有成果的研究领域,但在数据集和评估方面仍然面临重大障碍,需要手动标注术语,这是一项艰巨的任务,难度很大。术语和通用语言之间还缺乏清晰的区分,导致标注者之间的共识较少,增加了标注的歧义性。随着向机器学习和深度学习方法的不断发展,对带标注的数据集的需求变得越来越紧迫,不仅是为了评估,还因为“将机器学习或深度学习应用于 ATE 的主要问题之一是可靠的训练数据的可用性”。

通过对论文实验部分的阅读和总结,数据集主要分为公开数据集和基于特定研究的数据集。公开数据集是能够公开获取的带标注的数据集,具有广泛适用性,包括 GENIA、ACL RD-TEC、FAO 等,表 4 展示了常用数据集的统计信息。其中,GENIA 是评估 ATE 时最常用的数据集,用于生物医学文本挖掘的语义标注数据集;ACL 数据集是专门为 NLP 领域中的 ATE 评估而设计的,其假设是:拥有一个数据集,让 NLP 的研究人员可以自己成为领域专家,这将是一个巨大的优势。除了表中数据集外,还有一些较小的公开资源,如 TTCm 和 TTCw[2]。TTCw 语料库包含 103 篇关于风能领域的全文,TTCm 包含有关移动技术领域的 37 篇全文。

表 4　数据集摘要统计

数据集	领域	文本量(篇)	单词量(个)	术语量(个)	术语来源
GENIA	生物医学	2 000	494 000	35 104	手动标记
ACL	计算语言学	10 085	41 202 000	21 543	手动标记
ACL 2.0	计算语言学	300	33 000	3 095	手动标记
FAO	农业	779	26 672 000	1 554	作者的关键字
Europarl	政治	9 672	63 279 000	15 094	Eurovoc 词库

笔者发现公开数据集以英文为主,汉语研究主要以研究目的为导向,人工构建领域数据集。黄菡等[26]将裁判文书作为研究对象,从“中国裁判文书网”中抓取裁判文书 61 515 份,经过数据清洗后,人工标注了包括罪名、刑罚、法律原则、法律概念及法律条文 5 种类别的术语。为了抽取新能源汽车领域术语,马建红等[29]人工标注了专利文本 1 126 篇,并在 CAI 创新工具中得到验证。多语术语抽取研究是一个新兴领域,R. A. Terryn 等[32]收集了 3 种语言(英语、法语和荷兰语)和 4 个领域(腐败、盛装舞步、心力衰竭和风能)的语料,并设计了标注方案。基于特定研究的数据集涉及的领域范围广,还包括金融、军事、图书情报、科技文献、专利、网页文本等。

3.2.2 评价分析

ATE 评价的传统方法是与人工标注结果进行比较,并计算精度(实际的候选术语个数)、召回率(抽取出的正确术语个数)和 F 值(精度和召回率之间的调和平均值)。如黄菡等[26]利用准确率 P(Precision)、召回率 R(Recall)、F 值评价了法律术语识别的效果。这三个指标不能全面反映抽取结果的好坏,与噪声(错误提取的术语)和沉默(未提取出的术语)密切相关。此外,受试者工作特征曲线(receiver operating characteristic curve,ROC)也是一种评价方法,但在术语抽取领域不太常见。由于这些指标仅能衡量绩效,因此一些研究人员认为,更全面的评价协议是必要的。早在 1996 年,M. C. L Homme 等[33]广泛定义了 5 项预评价标准,以补充上述指标。在其他工作中,V. A. Sauron[34]提出了一种质量模型,该模型不仅计算精度或召回率,还可测量适用性、可靠性、可用性、可维护性和可移植性。同样只是使用 P、R、F,赵洪等[35]探讨了训练语料规模对抽取结果的影响,实验中计算了在 20%、40%、60% 和 80% 训练集比例下的抽取性能。D. Inkpen 等[36]考虑混合多种评价策略,并设计了用于促进 ATE 系统比较评价的工具。

3.3 应用分析

如表 5 所示,基于机器学习的术语抽取应用包括知识组织、自然语言处理、情报分析以及其他。在图情领域的应用主要体现在叙词表、本体等知识组织系统的构建、科技情报分析、专利术语抽取等内容,以支持情报系统的建设与服务。术语抽取是数据和知识获取的基本任务,也是许多复杂自然语言处理任务的预处理步骤,如信息检索、机器翻译、文本挖掘、关系抽取

等。其他应用指依据论文研究目标而进行的领域术语抽取任务，涉及的领域主要包括金融、军事、法律、医学、商业、农业。对应用情况进行总结有助于研究人员了解该领域的研究现状，指明研究方向，挖掘研究价值，探索未来的发展前景。以下从叙词表的维护更新、本体构建、自然语言处理以及情报分析 4 个方面详细介绍术语抽取的研究情况。

表 5 抽取技术的应用领域

主题	分类	关键词
抽取技术应用	知识组织	领域叙词表、地质学词典、本体学习、本体构建、学术资源本体、数字图书馆本体、Mesh、政务本体
	自然语言领域	信息检索、机器翻译、文本挖掘、问答系统、关系抽取、文本分类
	情报分析	专利术语抽取、科技情报分析、科技文献、科技政策、学术文献、信息技术
	其他	金融、军事、法律、医学、商业、农业

3.3.1 叙词表的维护更新

在生物医学、计算机科学、自然科学等领域，新术语会随着学科中新技术、新知识的产生而出现，为了促进领域叙词表资源的共享利用，叙词表的维护更新势在必行。M. Ikeda 等[37]从扩展多种叙词表的角度出发，利用机器学习进行候选术语的抽取，然后根据语法信息将相应领域的未注册术语加入到对应的叙词表。在科学计量学中使用叙词表和分类法来获取科学和技术信息一直受到关注，T. Kawamura 等[38]为了及时了解各种科学技术活动的最新趋势，提出利用 Word2Vec 工具从先进技术领域的文章摘要中获取领域相关的新概念和术语，以此来扩展领域叙词表。宋培彦等[39]研究了语义网环境下叙词表的构建方式，提出可以采用机器学习方法从语料库和文献资源中自动抽取术语，构造初始术语集。此外，还有像 Mesh 主题词表、Geo-Ref 地球科学叙词表等，在大数据背景下，为了提供全面的信息检索服务，基于机器学习的术语抽取将起着十分重要的作用。

3.3.2 构建领域本体

领域本体是共享概念模型的明确的形式化的规范说明，用公认的术语集合和术语之间的关系来反映该领域内的知识和知识结构，在语义信息交互、信息描述的规范化等方面起着重要作用。术语是领域本体构建的基本元素，术语抽取是本体学习中最基本也是至关重要的一步。为了提高题本构建的效率、降低本体构建的成本，B. Omelayenko[40]较早地利用机器学习的方法进行了术语提取、本体合并、更新以及实例的获取。李丽双[41]提出了基于条件随机场和主动学习相结合的领域术语抽取方法，实现了本体构建过程中一定程度的自动化，为制造企业知识管理的建模提供了较好的方法。为了构建领域学术本体，蒋婷[42]采用层叠条件随机场与 C-value 和规则相结合的方法分别对不同术语类型进行抽取。

3.3.3 自然语言处理

基于机器学习的术语抽取同样可以应用于自然语言处理领域，R. Gaizauskas 等[43]介绍了一种从 Web 源自动提取双语术语对的多组件系统 BiTES，首先自动从单语语料中提取术语，然后再从可比较的文档或平行语料中对齐提取的术语。G. Huang 等[44]发现网页上的括号里含有大量的术语翻译知识，为了提高抽取的召回率，作者提出了基于最大熵的术语识别系统 TermExt，并将抽取出的术语利用监督的机器学习方法进行机器翻译，实验表明，相比 baseline 抽取召回率提高了 11%。在信息检索领域，N. T. W. Khin 等[45]提出了基于 Web 查询分类算法的 IR 系统，系统包括领域术语提取、Web 查询分类和相关查询检索；一体化医学语言系统 UMLS 集成了 150 多部医学主题词表，广泛用于对互联网文献的检索和挖掘；IEEE 推出的顶层本体 SUMO 也试图将包括叙词表在内的知识组织工具进行融合，以提供更加全面的知识检索服务。

3.3.4 情报分析

在大数据环境下，通过情报分析进行科技信息监测和知识获取变得越来越重要，科技术语可以表征科技概念，表达科技数据的核心内容，是科技数据情报分析的重要内容之一。曾文等[46]介绍了基于深度学习算法的科技术语抽取方法，并在科技数据集上做出实验性的分析和结论；曾文、车尧等[47]以科技大数据为视角和分析对象，提出面向科技大数据情报分析服务的方法，并且设计研发了融合多种抽取算法的中文科技术语抽取方法，实验表明该方法在一定程度能够辅助情报研究人员进行数据的处理和分析。理论术语是大规模文献内容分析和跨学科知识转移深度揭示的基础，赵洪等[35]构建了面向理论术语抽取的深度学习模型。专利文献分析能够判断领域技术热点、预测技术发展趋势、帮助研发人员从中获得启发与借鉴，其中专利文献术语能够提供结构化知识，是专利文献分析的

关键组成部分[48-49]。

4 结语

本文采用文献计量分析法和内容分析法,在相关主题词下,对 Web of Science、CNKI 和维普数据库中与机器学习技术方法有关的论文进行了分析。通过文献计量对数据集的外部特征进行了宏观分析,包括年度趋势和核心机构,发现随着相关领域的快速发展术语抽取研究还处于上升期,可以通过关注“沈阳航空航天大学”“南京大学”等核心机构来进行学术追踪。之后,笔者重点对 169 篇中英文文献从抽取技术、数据集和评价以及应用3 个方面进行了主题分析,得到以下 3 点结论:

(1)统计机器学习方法的引入使术语抽取技术取得了很大的进步,但模型的识别性能很大程度上依赖标注语料的质量和特征工程。近几年,深度学习带来的新热潮又推动了术语抽取研究的发展,深度学习可以自动学习特征,减少了对领域知识的依赖。然而,从目前已有研究成果来看,术语抽取还远未得到解决,仍然是一个有挑战性的研究领域。大数据环境下,机器学习乃至深度学习将会是最有效的术语抽取方法。为了进一步提高模型的精度,值得考虑改进的地方还有很多,如采用混合方法、结合领域知识库、使用预训练模型等。

(2)自动术语抽取的数据集和评估方法对于量化最新技术的绩效至关重要,应包括文本语料库、黄金标准和评估指标。A. R. Terryn 等[32]在几种领域和语言中提供了一些标准数据集和标注策略。在 3.3.2 节中也介绍了国内外的一些数据集和评价方法,这些数据集对于正确评价术语抽取模型来说是无价的。在多源异构的数据环境下,数据集和评估方面仍然面临重大障碍,术语抽取理论体系需要完善,包括语料选取、评价指标和效果评价方法等。

(3)基于机器学习的术语抽取技术是知识系统、自然语言处理、情报分析等研究领域基础且重要的工作,具有较高的实用价值。应用不限于 3.3 节中指出的几方面,更多不同领域的应用还有待研究人员进一步探索。事实上,随着数据的海量化、异构化和复杂化,机器学习和深度学习会在术语抽取中起着越来越重要的作用。

本文仍有很多不足,例如,不能保证文献收集的全面性和准确性,在计量分析中存在误差,主题分析和应用领域分析的深度不够。研究希望尽可能准确地反映基于机器学习的术语抽取研究领域的现状,敬请广大专家学者批评指正。

参考文献:

[1] 术语工作原则与方法[J]. 术语标准化与信息技术,2003(1):45 – 48.

[2] ZHANG Z, GAO J, CIRAVEGNA F. Semre-rank: improving automatic term extraction by incorporating semantic relatedness with personalised pagerank [J]. ACM transactions on knowledge discovery from data, 2018, 12(5): 1 – 41.

[3] ASTRAKHANTSEV N. ATR4S: toolkit with state-of-the-art automatic terms recognition methods in scala [J]. Language resources and evaluation, 2018, 52(3): 853 – 872.

[4] CASTELLVí M T C, BAGOT R E, PALATRESI J V. Automatic term detection: a review of current systems [J]. Recent advances in computational terminology, 2001(2):53 – 88.

[5] MARSHALL P, BANDAR Z. Working towards connectionist modeling of term formation[C]//Proceedings of the international conference on computational intelligence. Heidelberg: Springer,1999: 522 – 529.

[6] BENGIO Y. A connectionist approach to speech recognition[J]. International journal of pattern recognition and artificial intelligence, 1993, 7(4): 647 – 667.

[7] 陈文亮, 朱靖波, 姚天顺, 等. 基于 Bootstrapping 的领域词汇自动获取[C]//全国第七届计算语言学联合学术会议论文集. 北京: 清华大学出版社, 2003: 67 – 72.

[8] KAUSHIK N, CHATTERJEE N. A practical approach for term and relationship extraction for automatic ontology creation from agricultural text[C]//Proceedings of the 2016 international conference on information technology. Bhubaneshwar:IEEE, 2016:241 – 247.

[9] STANKOVIC R, KRSTEV C, OBRADOVIC I, et al. Rule-based automatic multi-word term extraction and lemmatization[C]//Proceedings of the 10th international conference on language resources and evaluation. Portorož, Slovenia: European Language Resources Association, 2016: 507 – 514.

[10] DU L, LI X, LIN D. Chinese term extraction from Web pages based on expected point-wise mutual information[C]//Proceedings of the 2016 12th international conference on natural computation, fuzzy systems and knowledge discovery. Changsha:IEEE, 2016: 1647 – 1651.

[11] 李丽双,王意文,黄德根. 基于信息熵和词频分布变化的术语抽取研究[J]. 中文信息学报,2015,29(1):82 – 87.

[12] FRANTZI K T, ANANIADOU S, TSUJII J. The c-value/nc-value method of automatic recognition for multi-word terms[C]//Proceedings of the international conference on theory and practice of digital libraries. Berlin: Springer, 1998: 585 – 604.

[13] 周浪, 史树敏, 冯冲, 等. 基于多策略融合的中文术语抽取方法[J]. 情报学报, 2010 (3): 460 – 467.

[14] 王思丽,祝忠明,刘巍,等. 基于深度学习的领域本体概念自动

获取方法研究[J]. 情报理论与实践,2019(10):1 – 13.
[15] LOPEZ P, ROMARY L. HUMB: automatic key term extraction from scientific articles in GROBID[C]//Proceedings of the 5th international workshop on semantic evaluation. Los Angeles: Association for Computational Linguistics, 2010: 248 – 251.
[16] 赵欣. 基于最大熵的中文术语抽取系统的设计与实现[D]. 西安:西安电子科技大学,2012.
[17] SHIRAKAWA M, NAKAYAMA K, HARA T, et al. Wikipedia-based semantic similarity measurements for noisy short texts using extended naive bayes [J]. IEEE transactions on emerging topics in computing, 2015, 3(2): 205 – 219.
[18] ZENG W, LI X, LI H. Study on Chinese term extraction method based on machine learning[C]//Proceedings of the international conference of pioneering computer scientists, engineers and educators. Singapore: Springer, 2018: 128 – 135.
[19] PAN H S, ZHAO J Y. Combining syntactic information with HMM for term extraction[C]//Proceedings of the 2015 2nd international conference on information science and control engineering. Washington, DC: IEEE Computer Society, 2015: 170 – 173.
[20] 岑咏华, 韩哲, 季培培. 基于隐马尔科夫模型的中文术语识别研究[J]. 数据分析与知识发现, 2008, 24(12): 54 – 58.
[21] 章成志. 基于多层术语度的一体化术语抽取研究[J]. 情报学报, 2011, 30(3): 275 – 285.
[22] ZHENG D, ZHAO T, YANG J. Research on domain term extraction based on conditional random fields[C]//International conference on computer processing of oriental languages. Heidelberg: Springer, 2009: 290 – 296.
[23] ZHAN Q, WANG C. A hybrid strategy for Chinese domain-specific terminology extraction[C]//2015 11th international conference on semantics, knowledge and grids. Washington, DC: IEEE Computer Society, 2015: 217 – 221.
[24] RIGOUTS TERRYN A, DROUIN P, HOSTE V, et al. Analysing the Impact of supervised machine learning on automatic term extraction: HAMLET vs TermoStat[C]// Proceedings of the international conference on recent advances in natural language processing. Varna, Bulgaria: INCOMA Ltd. 2019: 1012 – 1021.
[25] CHI C Y, ZHANG Y. Information extraction from Chinese papers based on hidden markov model[J]. Advanced materials research, 2013, 846:1291 – 1294.
[26] 黄菡,王宏宇,王晓光. 结合主动学习的条件随机场模型用于法律术语的自动识别[J]. 数据分析与知识发现,2019,3(6):66 – 74.
[27] CHALAPATHY R, BORZESHI E Z, PICCARDI M. Bidirectional LSTM-CRF for clinical concept extraction[C]//Proceedings of the clinical natural language processing workshop. Osaka: The COLING 2016 Organizing Committee, 2016: 7 – 12.
[28] WANG R, LIU W, MCDONALD C. Featureless domain-specific term extraction with minimal labelled data[C]//Proceedings of the Australasian Language Technology Association workshop 2016. Australia: Australasian Language Technology Association, 2016: 103 – 112.
[29] 马建红, 张亚梅, 姚爽, 等. 基于 BLSTM_Attention_CRF 模型的新能源汽车领域术语抽取[J]. 计算机应用研究, 2019(5): 1 – 8.
[30] 刘宇飞,尹力,张凯,等. 基于深度迁移学习的技术术语识别——以数控系统领域为例[J]. 情报杂志,2019,38(10):168 – 175.
[31] ALFARONE D, DAVIS J. Unsupervised learning of an is-a taxonomy from a limited domain-specific corpus[C]//24th international joint conference on artificial intelligence. Buenos Aires: AAAI Press, 2015: 1434 – 1441.
[32] TERRYN A R, HOSTE V, LEFEVER E. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora[J]. Language resources and evaluation, 2019(6): 1 – 34.
[33] L'HOMME M-C, BENALI L, BERTRAND C, et al. Definition of an evaluation grid for term-extraction software [J]. Terminology international journal of theoretical and applied issues in specialized communication, 1996, 3(2): 291 – 312.
[34] SAURON V A. Tearing out the terms: evaluating terms extractors [C]//Proceedings of translating and the computer 2002. London: Aslib, 2002:1 – 18.
[35] 赵洪, 王芳. 理论术语抽取的深度学习模型及自训练算法研究[J]. 情报学报, 2018, 37(9): 923 – 938.
[36] INKPEN D, PARIBAKHT T S, FAEZ F, et al. Term evaluator: a tool for terminology annotation and evaluation [J]. International journal of computational linguistics and applications, 2016, 7(2): 145 – 165.
[37] IKEDA M, YAMAMOTO A. Extending various thesauri by finding synonym sets from a formal concept lattice [J]. Information and media technologies, 2017(12): 240 – 266.
[38] KAWAMURA T, KOZAKI K, KUSHIDA T, et al. Expanding science and technology thesauri from bibliographic datasets using word embedding[C]//2016 IEEE 28th international conference on tools with artificial intelligence. San Jose:IEEE, 2016: 857 – 864.
[39] 宋培彦,陈白雪,王星. 语义网环境下叙词表构建方法研究[J]. 情报科学, 2018, 36(2): 14 – 17.
[40] OMELAYENKO B. Learning of ontologies for the Web: the analysis of existent approaches [C]//Proceedings of the international workshop on Web dynamics. London: WebDyn@ ICDT. 2001: 16 – 25.
[41] 李丽双. 领域本体学习中术语及关系抽取方法的研究[D]. 大连:大连理工大学,2013.
[42] 蒋婷. 学科领域本体学习及学术资源语义标注研究[D]. 南京:南京大学,2017.
[43] GAIZAUSKAS R, PARAMITA M L, BARKER E, et al. Extracting bilingual terms from the Web[J]. Terminology international journal of theoretical and applied issues in specialized communica-

tion, 2015, 21(2): 205 –236.

[44] HUANG G, ZHANG J, ZHOU Y, et al. Learning from parenthetical sentences for term translation in machine translation[C]//Proceedings of the 9th SIGHAN workshop on Chinese language processing. Taipei: Association for Computational Linguistics. 2017: 37 –45.

[45] KHIN N T W, YEE N N. Query classification based information retrieval system[C]//2018 international conference on intelligent informatics and biomedical sciences. Bangkok:IEEE, 2018: 151 –156.

[46] 曾文,李辉,徐红姣,等. 深度学习技术在科技文献数据分析中的应用研究[J]. 情报理论与实践,2018,41(5):110 – 113.

[47] 曾文,车尧,张运良,等. 服务于科技大数据情报分析的方法及工具研究[J]. 情报科学,2019,37(4):92 –96.

[48] 俞琰,赵乃瑄. 融入术语知识的专利主题发现方法[J]. 图书情报工作,2018,62(21):118 – 126.

[49] 王健,殷旭,吕学强,等. 基于 CRFs 的专利文献领域术语抽取方法[J]. 计算机工程与设计,2019,40(1):279 – 284.

作者贡献说明:
邱科达:文献调研,论文撰写;
马建玲:选题建议,论文修改及润色。

A Statistical Analysis of Literature on Term Extraction Based on Machine Learning

Qiu Keda　Ma Jianling
Lanzhou Library Chinese Academy of Sciences, Lanzhou 730000
Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000
Department of Library, Information and Archives Management, School of Economics and Management,
University of Chinese Academy of Sciences, Beijing 100049

Abstract: [**Purpose/significance**] The purpose of this paper is to sort out and summarize the relevant content of the automatic term extraction research based on machine learning, and to provide a reference for related personnel in the field. [**Method/process**] Firstly, this paper applied literature measurement to conduct a macro analysis of the subject's annual trends and core institutions based on the analysis tools of CNKI and EndNote, then it carried out the subject analysis from 3 aspects: extraction of technical methods, data sets and evaluation, and application. [**Result/conclusion**] In recent years, term extraction research has made great progress, and is the basic work in the fields of knowledge systems, natural language processing, and information analysis. With the rapid development of natural language processing, extraction technology has begun to develop in the direction of deep learning, but the basic theoretical system of term extraction still needs to be improved, such as evaluation indicators, corpus selection and effect evaluation methods.

Keywords: term extraction　machine learning　knowledge organization　bibliometrics